LA-UR-79-837 CONF-790619--1

TITLE: A METHOD OF LINES APPROACH TO THE NUMERICAL SOLUTION OF CONSERVATION LAWS

AUTHOR(S): James M. Hyman

SUBMITTED TO: Third IMACS International Symposium On Computer Methods for Partial Differential Equation, June 20,21,22, 1979, Lehigh University, Bethlehem, Pennsylvania 18015.

University of California

LOS ALAMOS SCIENTIFIC LABORATORY
Post Office Box 1663 Los Alamos, New Mexico 87545
An Affirmative Action/Equal Opportunity Employer

fmB

A METHOD OF LINES APPROACH TO THE NUMERICAL SOLUTION OF CONSERVATION LAWS

by

James M. Hyman
Theoretical Division
Los Alamos Scientific Laboratory
Los Alamos, NM 87545

SUMMARY

New explicit finite difference methods are developed for approximating the discontinuous time dependent solutions of nonlinear hyperbolic conservation laws. The analysis is based on the method of lines approach of decoupling the space and time discretizations and analyzing each independently before combining them into a composite method. Particular attention is given analyzing to high order spatial differences, artificial dissipation and the accurate approximation of boundary conditions. Both a third order iterated leap-frog predictor-corrector and a second order iterated Runge-Kutta method are shown to have excellent stability and accuracy properties for the time integration. These methods are A-stable when iterated to convergence and have the special property of allowing for local improvements in the stability and accuracy of the computed solution.

15 References

The paper is designed to aid a scientist or engineer construct a numerical method specially taylored to a specific problem. The analysis requires an elementary knowledge of the numerical solution of ordinary differential equations, finite difference theory and gas dynamics.

CONTENTS

I. INTRODUCTION
II. CONSERVATION LAWS
   A. General Theory
   B. Euler Equations
III. SPACE DISCRETIZATION
   A. Phase and Damping Errors
   B. Grid Resolution
IV. BOUNDARY CONDITIONS
   A. Fictitious Points
   B. No Fictitious Points
   C. Global Relationships
   D. Imbedded Regions
   E. Characteristic Form
   G. Differential Form
V. ARTIFICIAL DISSIPATION
   A. Entropy Production
   B. Energy Cascade
   C. Gibb's Phenomenon
   D. Dispersion Error
   E. Stabilization of Time Integration Methods
   F. Differential Form
VI. TIME DISCRETIZATION
   A. Spectral Analysis
   B. Iterated Multistep Methods
   C. Stability
   D. Accuracy
VII. NUMERICAL ALGORITHMS
   A. Composite Methods
   B. Improved Artificial Dissipation
VIII. NUMERICAL RESULTS
   A. Riemann Problem
   B. Generalized Euler Equations
IX. SUMMARY, CONCLUSIONS AND RECOMMENDATIONS
REFERENCES

I. INTRODUCTION

The numerical solution of conservation laws is a highly complicated and problem-dependent process. The solution usually contains dynamic interactions between shock waves, rarefaction waves and contact discontinuities. A method developed for a particular test problem may or may not work for another with stronger (or weaker) shocks and contact discontinuities. Methods which work well in one space dimension may or may not be easily extended to two or three dimensions.

The analysis in this report is based on solving systems of conservation laws in one space dimension. We recognize that there exist many excellent methods to integrate systems of one-dimensional conservation laws. The addition of another numerical method designed only for these equations would not contribute significantly to the advancement of scientific computing — unless it generalizes to higher spatial dimensions. We also recognize that there is little hope of generalizing a method which fails in one dimension to one that works in higher dimensions.

The approach is based on the philosophy of the method of lines (MOL). (Hyman (1976)). In the method of lines, the space and time discretizations of a partial differential equation (PDE) are decoupled and analyzed independently. First a method is selected to discretize the differential equation in space and incorporate the boundary conditions. The spectrum of this discrete operator is then used as a guide to choose an appropriate method to integrate the equations through time.

The dissipative effects of a numerical method are crucial to constructing reliable methods for conservation laws. This is particularly true when the solution is discontinuous as in a shock wave or contact discontinuity. (All variables are discontinuous across a shock wave, while across a contact discontinuity the density is discontinuous and the pressure and velocity are continuous). The numerical dissipation can be shaped or controlled by adding an artificial dissipation term explicitly to the differential equation. This dissipation is the leading truncation error in the numerical approximation and is chosen on the basis of the expected form of the solution. For example, more dissipation is added to calculations where there are strong shock wave interactions than when the important aspect of the calculation is to determine the location of a contact discontinuity.

Choosing an accurate method to accomplish each of these tasks, space and time discretization and incorporating artificial dissipation in the numerical solution, determines the success of the calculation. In Sections III through VI we will consider each choice independently and combine them in Section VII to develop a class of particularly good explicit finite difference methods. In Section VIII, we present some numerical examples to illustrate the properties of the different methods and analyze their results in a summary, Section IX.

Before developing a good method for any system of PDEs one must have a basic understanding of the equations being solved. We now give a brief review of the general theory of conservation laws.

II. CONSERVATION LAWS

General Theory

A vector quantity W of length N is conserved as it evolves under the flow of a conservation law if the amount of each substance W_j, $j=1,2,\ldots,N$, contained in any fixed volume V is due entirely to the flux F_j across the boundary ∂V of V. These conservation laws can be expressed in integral form as

$$\frac{d}{dt}\int_V W_j\,dx = -\int_{\partial V} F_j(W)\cdot n\,dS \tag{2.1}$$

where n denotes the outward normal to the boundary.

Moving the time derivative under the integral sign and applying the divergence theorem Eq. (2.1) can be rewritten as

$$\int_V \frac{\partial}{\partial t} W_j + \operatorname{div} F_j(W)\,dx = 0\ . \tag{2.2}$$

By letting the volume V shrink to a point we obtain the system of PDEs

$$\frac{\partial}{\partial t} W_j + \operatorname{div} F_j(W) = 0,\ j = 1,2,\ldots,N \tag{2.3}$$

at every point where W and F are differentiable.

In one space dimension Eq. (2.3) can be written in vector form as

$$W_t + F(W)_x = 0 \tag{2.4}$$

or

$$W_t + G(W)W_x = 0 \tag{2.5}$$

where G is the N by N matrix gradient of F with respect to W.

Equation (2.5) is a first-order quasi-linear system of PDEs. This system is hyperbolic and well-posed if the eigenvalues of G are distinct and real. These eigenvalues, called the characteristic velocities, are the local signal speeds at which sharp disturbances propogate.

It is well known that a system of nonlinear conservation laws may fail to have a continuous solution after a finite time. Since conservation laws are derived from integral relations (2.1) these generalized solutions may still be admitted as long as they are measurable and bounded. There are instances, as in the Euler equations of gas dynamics, that there may be many different generalized solutions satisfing Eq. (2.1) with the same initial data. Within this set only one of these solutions has any physical significance. An important consideration in constructing a numerical method is to build a mechanism into the difference scheme that will automatically choose the physically relevant solution.

The physically relevant solution must satisfy the differential equation (2.4) in smooth regions and fulfill two additional constitutive relations across any discontinuities in the flow. The first constitutive relation, called the Rankine-Hugoniot jump conditions, states that the discontinuity must propagate with speed s satisfing the jump conditions

$$s[W_j] = [F_j(W)]\ ,\ j = 1,2,\ldots,N\ .$$

Here [ ] denotes the jump of the quantity in brackets across the discontinuity. These jump conditions are satisfied by the numerical solution if the equations are solved in divergence form (Eq. (2.4) and the flux function is differenced with centered differences.

The second constitutive relation, called the entropy condition, states that entropy must increase across the shock discontinuity. This condition is satisfied by the limiting solution of the viscous equations as the viscosity is decreased to zero. Numerical methods for solving discontinuous solutions of Eq. (2.4) have a small artificial viscosity that helps select the physically relevant solution.

The methods developed in this paper will be described in terms of the Euler equations of gas dynamics. However, most of the techniques and results are equally valid for other hyperbolic systems. For a further discussion of the mathematical theory of general hyperbolic systems of conservation laws we refer the reader to Lax (1973) and (1976).

Euler Equations

The one-dimensional Eulerian equations of gas dynamics can be written in divergence form as

$$W_t + F(W)_x = 0\ , \tag{2.6}$$

$$W = \begin{pmatrix} \rho \\ m \\ E \end{pmatrix},\ F(W) = uW + \begin{pmatrix} 0 \\ p \\ pu \end{pmatrix}\ .$$

where ρ = mass density, u = velocity, $m = \rho u$ = momentum, $E = \rho(I+\frac{1}{2}u^2)$ = total energy per unit volume, I = internal energy and p = pressure.

Equation (2.6) is hyperbolic if pressure is an increasing function of density at constant entropy. This is the case if we assume the equation of state to be that of a polytropic gas, i.e. $p = (\gamma-1)I\rho$. The parameter γ is a constant greater than one and equal to the ratio of the specific heats of the gas. For this equation of state we have

$$\frac{dp}{d\rho} = \frac{\gamma p}{\rho} = c^2 > 0$$

at constant entropy. The quantity c is called the local sound speed of the gas and is related to the characteristic velocities u, $u+c$ and $u-c$ of Eq. (2.6).

III. SPACE DISCRETIZATION

To solve Eq. (2.6) numerically we must first chose an appropriate approximation of the spatial derivatives. The guiding principle in choosing a spatial approximation is that <u>the discrete model should retain as closely as possible all the crucial properties of the original differential equation</u>. Equations (2.6) reflect principles of conservation of mass, momentum, and energy which are the basis for the mathematical theory of fluid dynamics. These properties should be preserved in the difference formulation. This is best accomplished if the equations are integrated and differenced in divergence form using centered finite differences.

Phase and Damping Errors

The derivative of the flux function F determines the phase velocities of the solution and hence the shock speeds. Therefore, the errors in approximating F and its derivative should be made as small as possible. These errors can be divided into two classes; phase or dispersion errors and damping or dissipation errors.

The analysis is based on computing the errors for numerical approximations of traveling wave solutions to Eq. (2.6). Consider the solution to Eq. (2.6) with periodic boundary conditions on the unit interval and constant initial pressure and velocity v. The solution is a traveling wave and satisfies the simple linear hyperbolic convective equation

$$\rho_t + v\rho_x = 0 , \tag{3.1}$$

with the initial conditions $\rho(x,0) = g(x)$, and the solution $\rho(x,t) = g(x-vt)$.

Let N be a natural number, define $\Delta x = 1/(2N+1)$ and $x_k = k\Delta x$, where $k = 0,1,\ldots,2N$. At these points $g(x)$ can be represented by the finite Fourier series

$$g(x_k) = \sum_{j=-N}^{N} \hat{g}_j \exp(2\pi i j x_k)$$

where

$$\hat{g}_j = \frac{1}{2N+1} \sum_{k=0}^{2N} g(x_k) \exp(-2\pi i j x_k) .$$

Then

$$\rho(x,t) = \sum_{j=-N}^{N} \hat{g}_j \exp(2\pi i j(x-vt)) \tag{3.2}$$

is a solution of (3.1) which takes on the initial values at the grid points x_k.

To approximate (3.1), we begin by replacing the spatial derivatives with second order centered differences. This results in a system of ordinary differential equations

$$R_t + HR = 0 \tag{3.3}$$

$$R = \begin{bmatrix} \rho(x_0,t) \\ \vdots \\ \rho(x_{2N},t) \end{bmatrix} , \quad H = \frac{v}{2\Delta x} \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 & -1 \\ -1 & 0 & 1 & 0 & & & 0 \\ 0 & -1 & 0 & 1 & & & \cdot \\ \vdots & & & & \ddots & & \vdots \\ 0 & & & & -1 & 0 & 1 \\ 1 & 0 & \cdots & & & -1 & 0 \end{bmatrix}$$

Equation (3.3) can also be solved exactly and the solution is

$$\rho_k(t) = \sum_{j=-N}^{N} \hat{g}_j \exp(2\pi i j(x_k - v_j t)) , \tag{3.4}$$

where

$$v_j = v \frac{\sin 2\pi j\Delta x}{2\pi j\Delta x} .$$

The eigenfunctions of the approximate solution (3.4) are the same as in the exact solution (3.2) but the temporal eigenvalues differ. The eigenvalues of H are purely imaginary, as they should be for a hyperbolic operator, and so all the errors occur in the phase of the approximate solution.

The principle of linear superposition states that a general solution to Eq. (3.1) can be expressed as a sum of solutions to Eq. (3.1), each of which consists of only a single frequency. By analyzing the error in each frequency of the solution we can compare the accuracy of different numerical methods.

The phase error in the k-th mode is

$$e_2(k) = v2\pi kt \left(1 - \frac{\sin 2\pi k\Delta x}{2\pi k\Delta x}\right) \cong \frac{vt}{6h} (2\pi k\Delta x)^2 . \tag{3.5}$$

We define $M_2 = (k\Delta x)^{-1}$ as the number of mesh points per wave length. The error can be expressed as

$$e_2(k) \cong \frac{(2\pi)^3}{6M_2^2} vkt .$$

Observe that the phase and maximum errors of the approximate solution are related if the initial values consist of a single frequency $g(x) = \exp(2\pi i k x)$, then the maximum error of the approximate solution at the mesh points x_j is

$$\| \rho(x_j,t) - R_j(t) \|$$

$$= \| \exp(2\pi i k(x_j - vt)) - \exp(2\pi i k(x_j - v_k t)) \|_\infty$$

$$= \| \exp(2\pi i k x_j)(\exp(-2\pi i k v t) - \exp(-2\pi i k v_k t)) \|_\infty$$

$$\cong \| \exp(2\pi i k x_j)(v - v_k)\, 2\pi i k t \|_\infty$$

$$\cong e_2(k)$$

<u>Grid Resolution</u>

The number of mesh points needed per wave length for a given phase or maximum error e is

$$M_2 \cong 2\pi \left(\frac{\pi}{3e} vkt\right)^{\frac{1}{2}} .$$

The number of mesh points M_2 is dependent not only on the error but also on the square root of the computation time interval, the wave speed v, and the relevant modes of the initial data.

If fourth order centered differences

$$(\rho_i)_x \cong (-\rho_{i+2} + 8\rho_{i+1} - 8\rho_{i-1} + \rho_{i-2})/(12\Delta x)$$

are used the phase error in the k-th mode is

$$e_4(k) = v\, 2\pi\, kt \left(1 - \frac{8 \sin 2\pi k\Delta x - \sin 4\pi k\Delta x}{12\pi\, k\Delta x}\right)$$

$$\cong \frac{(2\pi)^5}{30M_4^4} vkt .$$

The number of mesh points needed per wave length for a given error e is

$$M_4 \cong 2\pi \left(\frac{\pi}{15e} vkt\right)^{\frac{1}{4}} .$$

The corresponding relationships for sixth order centered difference

$$(\rho_i)_x \cong (\rho_{i+3} - 9\rho_{i+2} + 45\rho_{i+1} - 45\rho_{i-1} + 9\rho_{i-2} - \rho_{i-3})/(60\Delta x)$$

are

$$e_6(k) \cong \frac{(2\pi)^7}{140\, M_6^6} vkt ,$$

and

$$M_6 \cong 2\pi \left(\frac{\pi}{70e} vkt\right)^{1/6} .$$

When the initial values consist of a single mode then for a given phase or maximum error the relationship between the number of mesh points needed per wave length is

$$M_2 \cong \frac{\sqrt{5}}{2\pi} M_4^2 \cong \frac{1}{(2\pi)^2} \left(\frac{70}{3}\right)^{\frac{1}{2}} M_6^3$$

or

$$M_2 \cong 0.36 M_4^2 \cong 0.12 M_6^3 .$$

This relationship is independent of the frequency, the phase error and the computation time. The table below compares the number of points per wave length necessary to obtain a given accuracy using second, fourth and sixth order centered differences.

2nd order M_2	4th order M_4	6th order M_6	Accuracy $e/(vkt)$
4	4	3	2.6
8	5	4	0.65
16	7	5	0.16
32	10	7	0.04
64	14	8	0.01
128	19	10	0.0025
256	27	13	0.0006

Table I. Points per wavelength for second, fourth and sixth order difference to have the same accuracy.

These error approximations have been verified numerically for accuracies $e/(vk) \geq 0.0001$ by integrating the discrete equations very accurately through time.

In a calculation where the solution contains many different frequencies, the high modes (2-5 points per wavelength) are approximated equally poorly with all the methods. The middle modes (6-16 points per wavelength) are computed much more accurately with the fourth and sixth order differences than with the second order method. The sixth order differences are more accurate for the lower modes than either second or fourth order differences.

The relationship of the accuracies of the different methods compared to the number of points per wavelength is even more impressive in higher dimensions. In two space dimensions the numbers in Table I should be squared; in three dimensions cubed.

There are disadvantages in using the higher order differences. They require more work to evaluate, there is the added complexity of adding two or three fictitious points at the boundaries, and the differential difference equations (3.3) are stiffer, resulting in more stringent time step restrictions. (See Section VI).

The enormous increase in accuracy outweighs all of these disadvantages. The difference formulas are more costly to evaluate, but most of the computer time is spent evaluating the flux functions, not differencing them. The actual increase in computer time is only a few percent when the same number of mesh points are used in both a second and a fourth order calculation.

The additional complexity or storage at the boundary cannot be avoided, but it should not deter the use of high order difference in the interior of the region of integration. It is desirable but not essential that the boundary conditions are approximated to the same order of accuracy as used in the interior. Numerical experiments have shown that the overall quality of the computation increases when fourth or sixth order differences are used in the interior instead of second order differences even when the boundary approximation remains at only second order in both calculations.

The other disadvantage of high order differences is they decrease the upper bound on the time step when compared to second order differences. The stability bound is reduced by 3/4 for fourth order differences and 6/11 for sixth order. This is not as severe a restriction as it might seem. The stability bounds of the iterated multi-step methods described in Section VI are greater than standard explicit methods. Also, the stability limit is proportional to $1/\Delta x$. When using high order differences you need fewer points to approximate the solution with the same accuracy than when using lower order differences. When the number of mesh points is reduced, the upper stability bound on the time step is increased. Therefore the overall work in a high order approximation on a course mesh may be much less than a lower order calculation with the same accuracy on a finer mesh.

The next step is to determine if this linear analysis is applicable to nonlinear equations with shocks and contact discontinuities. Figure 1 displays the solution to two initial value problems for Eq. (2.6). Each problem was solved twice, the only difference being that the spatial differences were varied from second to fourth order.

The exact traveling wave solution (dashed line) and computed solution (solid line) to Eq. (2.6) with v=1 are shown in Fig. 1 at time t=1. In this calculation we used eight mesh points per wavelength.

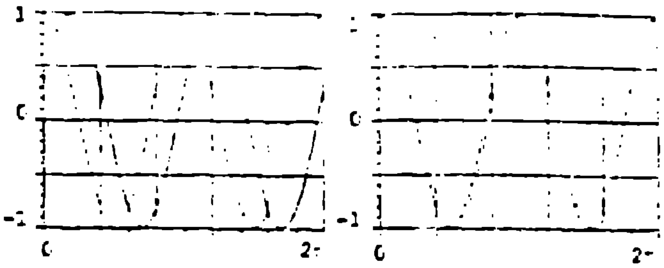

Fig. 1a 2-nd order Fig. 1b 4-th order

Notice the dramatic reduction in phase error between the second and fourth order differences.

The second example is the shock tube Riemann problem described completely in Section VIII. The shock and contact discontinuity in Figs. 1c and 1d are much sharper when the higher order differences are used. Also, the post shock oscillations are reduced even though the same artificial dissipation was used in each calculation. The high order differences are able to resolve the discontinuities better and require less artificial dissipation to eliminate post shock oscillations.

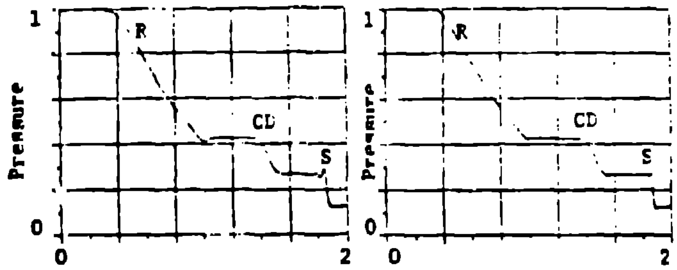

Fig. 1c 2-nd order Fig. 1d 4-th order

IV. BOUNDARY CONDITIONS

Before calculating the solution to any differential equation one should determine if the boundary conditions are consistent with a well posed problem. A numerical method cannot be expected to generate reasonable results for a problem which does not have a well defined reasonable solution. The importance of proper boundary conditions cannot be overstressed, the boundary conditions exert one of the strongest influence on the behavior of the solution. Also, the errors introduced into the calculation from improper boundary conditions persist even as the mesh spacing tends to zero.

A common error in prescribing boundary conditions for hyperbolic equations, such as the Euler equations, is to over or under specify the number of boundary conditions. Overspecification usually results in nonsmooth solutions with mesh oscillations near the boundary. Underspecification does not insure the solution is unique and the numerical solution may tend to wander around in steady state calculations. In either case the results of the calculation are not accurate and one should be skeptical of even the qualitative behavior of the solution.

When incorporating the boundary conditions into the discrete equations the spectrum of the discrete operator should be perturbed as little as possible from the spectrum of the differential operator. This can best be done by enforcing constituent relationships on the difference equations such that the discrete equations are consistent with as many relationships that can be derived from the boundary conditions and differential equation as possible.

Fictitious Points

To define the solution at fictitious points the boundary conditions are differentiated with respect to time and
obtain differential constraints for the extrapolation formulas. This technique will be shown for reflecting boundary conditions to the Euler equations.

The reflecting boundary conditions for a thermally insulated wall for Eq. (2.6) at $x = x_o$ are

$$u(x_o,t) = 0, \quad I_x(x_o,t) = 0. \tag{4.1}$$

The thermally insulating boundary condition, $I_x = 0$, is obtained from the limit of the viscous dissipative equations as the viscosity and heat dissipation tend to zero. This condition is necessary to prevent a boundary layer in the difference approximation of inviscid calculations due to the presence of artificial dissipation.

To incorporate these boundary conditions into our numerical solution when using fourth-order centered differences we will introduce two fictitious points at $x_{-1} = x_o - \Delta x$ and $x_{-2} = x_o - 2\Delta x$ outside the region of integration. At these points we need an approximation to ρ, ρu, and E to preferably fourth-order.

Combining Eqs. (2.6), and (4.1) at $x = x_o$ we have

$$0 = (\rho u) = -(\rho u)_t = (\rho u^2 + p)_x = p_x = (\gamma - 1)\, I\rho_x,$$

$$E_x = [\rho(I + \tfrac{1}{2} u^2)]_x = 0,$$

and

$$(\rho u)_{xx} = -(\rho_t)_x = -(\rho_x)_t = 0.$$

Since these equations are valid for all time and $\gamma \neq 1$ we have

$$\rho_x = m = m_{xx} = E_x = 0 \tag{4.2}$$

as auxiliary boundary conditions at $x = x_o$ consistent with the original problem. This procedure can be continued to give

$$\rho_{xxx} = m_{xxxx} = E_{xxx} = 0. \tag{4.3}$$

The nonphysical solution at the fictitious points outside the region of integration needs to be chosen such that a finite difference approximation of Eqs. (4.2) and (4.3) are satisfied at the boundary.

When we replace the derivatives in the auxiliary boundary conditions by the standard centered five point finite differences we see that Eqs. (4.2) and (4.3) are satisfied if and only if

$$\rho_{-i} = \rho_i,\; m_{-i} = -m_i,\; E_{-i} = E_i \tag{4.4}$$

for $i = 1$ or 2. Thus reflecting the solution symmetrically or antisymmetrically as in Eqns. (4.4) may appear to be only first-order but in fact is a very accurate approximation of the boundary conditions for the thermally isolated boundary given by Eq. (4.1).

No Fictitious Points

There is not always a simple extrapolation formula such as Eq. (4.4) to extend the solution to the fictitious points. For these problems it is often better to use uncentered differences near the boundary. This method will be described for the linear hyperbolic system of M equations

$$W_t = H(x)W_x \tag{4.5}$$

with the boundary conditions

$$SW_o = b(t),\; x = x_o. \tag{4.6}$$

Difficulties arise in defining the solution at the boundary when $0 < \text{Rank}(S) < \text{Rank}(H) = M$. If $\text{Rank}(S) = 0$ then all the characteristics are outgoing and using either uncentered differences at the points near the boundary or straight forward extrapolation to the fictitious points gives accurate results. When $\text{Rank}(S) = M$ then all the characteristics are entering the boundary and all the components of the solution can be solved for on the boundary. Uncentered spatial differences can then be used at the points near the boundary will result in an accurate approximation of the boundary conditions.

When $\text{Rank}(S)$ is greater than zero but less than M then by differentiating Eq. (4.6) with respect to time and replacing W_t from Eq. (4.5) we have

$$SH(x)W_x = b'(t),\; x = x_0. \tag{4.7}$$

Approximating W_x by second-order one-sided differences gives us

$$SH_o W_o = [SH_o(4W_1 - W_2) - 2\Delta x b'(t)]/3 \tag{4.8}$$

where $H_o = H(x_o)$. Equation (4.8) gives us additional information about the boundary conditions that is consistent with both the original boundary condition (4.6) and the differential Eq. (4.5). If

$$\text{Rank}\begin{pmatrix} S \\ SH_0 \end{pmatrix} < M,$$

we still do not have enough boundary conditions to solve for W_0 uniquely. By differentiating Eq. (4.7) with respect to time,

$$SH(W_t)_x = SH(HW_x)_x = b''(t),\; x = x, \tag{4.9}$$

and replacing the spatial derivative with finite differences we have

$$SH_0(H_0+H_1)\, W_0 = SH_0[H_0+2H_1+H_2)\, W_1 - (H_1+H_2)W_2] + 2\Delta^2 b''(t) = O(\Delta x^3)\,. \tag{4.10}$$

to give us an additional relationship.

It is often the case that H_0 is nonlinear and the above procedure must be iterated. Usually one or two iterations are sufficient for a stable accurate boundary approximation.

Once W_0 has been found we can use uncentered finite differences to approximate the spatial derivaties at the mesh point nearest the boundary or we can extrapolate the solution to fictitious points outside the region of integration. This extrapolation can be done by replaceing the derivatives in Eqs. (4.7) and (4.9) with second order centered differences and solve for W_{-1}.

Global Relationships

All of the boundary conditions considered so far have been local. That is, they depend only on the value of the solution and its derivative at the boundary, independent of their values in the interior. Some of these boundary conditions are derived from a global relationship. For example, the reflecting boundary condition (u=0) is related to the conservation of mass,

$$\frac{d}{dt}\int_0^1 \rho dx = -\int_0^1 (\rho u)_x dx = (\rho u)_0 - (\rho u)_1 = 0$$

Suppose the solution satisfies the functional global relationships

$$h_j(W) = 0 \quad , \quad j = 1,2,\ldots K \tag{4.11}$$

such as the conservation laws for mass, momentum and energy,

$$h_j(W) = \int_0^{b(t)} W_j\, dx + R_j(t) = 0\,. \tag{4.11}$$

These functional relationships can be approximated by their discrete analog equations

$$H_j(W) = 0 \quad , \quad j = 0,1,2,\ldots k \tag{4.12}$$

where the discrete functional H_j approximates the h_j with at least the same order of accuracy as the finite difference approximation to the Euler equations. The numerical approximation of the local form of the boundary conditions is not always sufficient to guarantee the global conservation laws they are derived from. In these situations (usually arising when there are moving boundaries or boundary layers) the global relation should be incorporated directly into the difference equations. This can best be done using a numerical technique based on the work of Isaacson (1977).

Since we require the modified solution to be consistent with the differential equation, one of our functional relationships must reflect this consistency. For example, the weak form (See Lax (1973)) of the Euler equations can be written as

$$h_0(W) = \int (W_t+F_x)\phi = 0$$

where ϕ is an arbitrary test function. This test function can be chosen to vanish outside the regions where the solution is to be varied. The time derivative can be replaced by finite differences such as the trapizodial rule

$$h_0(W) \equiv \int [W^{n+1}-W^n+ \frac{\Delta t}{2} (F_x^{n+1}+F_x^n)]\phi dx = 0$$

or some other implicit difference formula using the solution at $t = (n+1)\Delta t$. The spatial derivatives and integration are both approximated by finite differences.

Once the solution satisfying (4.12) has been advanced from $t = n\Delta t$ to $t = (n+1)\Delta t$ by any numerical method there is a residual error in the functional relationship

$$H_j(W_{n+1}^{(0)}) = \epsilon_0 \ .$$

The solution at time $(n+1)\Delta t$ is then modified by Newtons method to reduce this residual error.

In an actual calculation one rarely needs to vary the solution at all the mesh points and restricts the iteration to varying the solution at the points in a neighborhood of the particular problem causing loss of conservation, such as a moving boundary. In one dimension this usually means only one or two points need to be modified, in higher dimensions the iteration may include one or two lines or planes of points.

Imbedded Regions

There are many initial boundary value problems where it is essential to introduce artificial boundaries to reduce the computing time and storage of a calculation. These problems are usually posed in a domain much larger than the subregion where the solution is of interest. The subregion is blocked off and imbedded in the original problem by creating artificial boundaries. The boundary conditions at the artificial boundary are chosen such that the solution on the full domain would automatically satisfy these internal boundary conditions if the full problem were solved. The goal, of course, is to approximate the original problem as closely as possible on the reduced domains.

The boundary conditions must be consistent with a well posed initial boundary value problem in the reduced domain. Overspecifying the boundary conditions insures giving the wrong answer and underspecifying them does not guarantee a unique solution. Thus one must be careful to prescribe the correct number of boundary conditions according to a linearized analysis of the incoming and outgoing characteristics at the artificial boundary. These problems cannot be swept away by using one sided differences and including extra artificial dissipation to stablize the results. Invariably, when this is done, the numerical solution in subsonic flow problems will not be accurate.

We will now describe a computationally efficient approach to incorporate an artificial boundary into a flow which avoids both of these constraints. The method generalizes easily to higher dimensions and to systems other than the Euler equations.

Consider the initial boundary value problem for the Euler equations on the half line $[0,\infty)$, with reflecting boundary conditions at $x=0$, suppose we are interested only in the behavior of the solution in the interval $[0,1]$ and wish to restrict the domain of our computation to a neighborhood of this region. First we map the the interval $[1,\infty)$ into $[1,b)$ with a map such as

$$y = \begin{cases} x & 0 \le x \le 1 \\ b+(1-b)/x & 1 < x < b \end{cases} \qquad (4.15)$$

In this new coordinate system Eq. (2.6) transforms to

$$W_t + s(y)\ F_y = 0\ , \quad y\epsilon[0,b)$$

where

$$s(y) = \begin{cases} 1 & 0 \le y \le 1 \\ (b-y)/(b-1)^2 & 1 < y < b \end{cases} \qquad (4.16)$$

The solution to (4.16) is identical to the solution of our original problem. Therefore, the transformed system has the correct number of signals entering and leaving through the artificial break point at $x=1$.

The equally spaced mesh on $[0,b)$ corresponds to a variable mesh in the original coordinate system. The variable mesh spacing is constant in $[0,1]$ and increases in $(1,\infty)$.

In this transformed system a wave slows down in the region $(1,b)$ and approaches zero speed as x nears b. This causes a wave train to squeeze up, with the lower frequencies being pushed into higher ones as in Fig. 2a below. These high frequencies cannot be computed accurately and it is best to add some dissipation to damp them out as they approach the transformed boundary b. This damping should be chosen such that the signals propagating into the region of interest $[0,1]$ depend in some sense on an average of the solution outside this region, i.e. $(1,b)$. A possible form for the dissipation is

$$W_t + s(y)F_y = (\Delta y d(y)W_y)_y \qquad (4.17)$$

where

$$d(y) = \begin{cases} 0 & 0 \le y \le 1 \\ \delta[(y-1)/(b-1)]^2 & 1 < y \le b. \end{cases}$$

The graph in Fig. 2b shows the functional form of the two coefficients. Notice that the equation is unchanged in the interval $[0,1]$ and becomes parabolic in the interval $(1,b]$. In fact at $y=b$ the equation reduces to a simple diffusion equation.

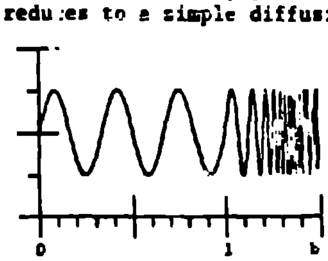

Fig. 2a Graph of $\sin(4\pi x(y))$

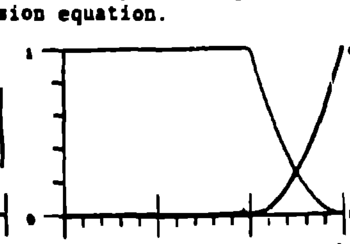

Fig. 2b Form of d and s

Boundary conditions must be given for all the variables at $y=b$ for the problem to be well-posed. The boundary condition for steady flow at infinity ($W_y=0$) gave the best results in a series of test problems.

By imbedding the equation in the subregion into a well-posed problem in a slightly larger domain the difficulty of maintaining the correct number of boundary conditions at the artificial boundary was assured automatically. Furthermore, the information entering the region at this boundary depends on some global average properties of the solution outside the subregion.

The number of points outside the imbedded subregion for the approximation to be accurate and prevent superfluous reflections depends on the strength of the outgoing waves and their angle of incidence to the boundary. The stronger the wave or the smaller the angle of incidence the more points are needed.

In one dimension the angle of incidence is always 90° and very few extra points are needed. In two and three dimensions three to five points may be needed for a strong shock wave glancing the boundary.

Characteristic Form

In problems where the solution is sensitive to the approximation of the boundary conditions it may be more stable to transform the boundary conditions or the equation into characteristic form. The extrapolation formulas are then derived to extrapolate the outgoing characteristic variables to the fictitious points. These formulas, as before, should incorporate as many relationships that can be derived from the boundary conditions and differential equation as possible.

At a subsonic inflow boundary the boundary conditions should be of the form

$$u = \alpha_1(u-c) + \beta_1(t) \qquad (4.18)$$

and

$$u+c = \alpha_2(u-c) + \beta_2(t) \ , \qquad (4.19)$$

where α_i and β_i are functions of t alone. At subsonic outflow boundary the boundary conditions should be of the form

$$u-c = \alpha_3 u + \alpha_4(u+c) + \beta_3(t) \ . \qquad (4.20)$$

When the flow is supersonic at the boundary then either three or no boundary conditions are given, depending on whether it is an inflow or outflow boundary.

Characteristic variables are also important when no amount of algebra seems to yield enough relationships to uniquely define all the solution variables at the fictitious points. When this happens one is forced to extrapolate on some of the variables without any boundary relationships to guide the extrapolation. It is usually best to extrapolate on outgoing characteristic variables and use their values at the fictitious points to provide the extra needed information.

Differential Form

Whatever extrapolation formula is used there may be some inherent truncation error in the extrapolated solution at the fictitious points. Some of these truncation errors can be eliminated by changing the differential form of the equation at the boundary.

For example, the reflecting boundary conditions (4.1) and (4.2) can be incorporated in the Euler equations at the boundary to give

$$\begin{pmatrix} \rho \\ m \\ E \end{pmatrix}_t + \begin{pmatrix} m \\ 0 \\ pu \end{pmatrix}_x = 0 \qquad (4.21)$$

at the boundary. By differencing these equations, rather then Eq. (2.6), at the boundary we have prevented some of the possible truncation errors inherent in the extrapolation formula, from creeping into our calculation. This is true even when the extrapolation formulas are based on a finite difference approximation of the boundary condition differential relationships.

Notice that the modified Eq. (4.21) has been kept in divergence form. This is particularly important to maintain conservation when shocks are reflected at the wall.

Using the modified differential form of the equations is especially important when there is a removable singularity at the boundary. These terms should be replaced by their equivalent form obtained using L'hopital's rule.

V. ARTIFICIAL DISSIPATION

Artificial dissipation or artificial viscosity is a special form of truncation error either inherent to a finite difference approximation or resulting from explicitly adding an additional term to the equation. The purpose of the artificial term is to remove many of the numerical difficulties by dissipating or damping out the high frequencies of the solution. This approach does in some sense mock up the effects of the viscous and dissipative terms discarded in the derivation of the Euler equations in that it primarily dissipates the high wave numbers, but it has little to do with true heat dissipation or viscosity.

There are five primary reasons for including artificial dissipation in the numerical approximation. They are:

1. To achieve proper entropy production across shock fronts.
2. To solve the problem of the energy cascade when computing only a finite number of modes.
3. To compensate for spatial errors, such as the Gibb's phenomenon, near discontinuities in the solution.
4. To compensate for the dispersion error in the numerical scheme.
5. To stabilize certain time differencing methods.

The form of a good artificial dissipation term tailored for a particular problem will depend on which of these points are most important. It is therefore essential to designing a numerical method to have a basic understanding of each of them. In this section we will review each reason for adding artificial dissipation and suggest a form which will, with luck, work for a large class of problems.

Entropy Production

The most common reason given for adding artificial dissipation is so that one can calculate shock waves. Entropy increases across a shock front, but Eq. (2.6) has no mechanism for the increase. We must add a term to the equation which will allow entropy to increase by the proper amount. The term should be in conservation form to maintain the Rankine-Hugoniot jump conditions and therefore give the correct shock speed.

Another desired effect of the artificial dissipation is to smooth out nonphysical discontinuities in the flow. That is, it would be advantageous if the artificial dissipation were formulated in such a way that physical shocks are stable and nonphysical sudden compression shocks are unstable.

It is important to add enough artificial dissipation to elliminate numerical oscillations around the shocks. These oscillations can destroy the accuracy of the calculation by creating nonlinear instabilities or introducing nonphysical features in the flow such as negative mass or pressure. The oscillations may generate new artifacts into the calculation such that the numerical calculation is stable but converges to the wrong solution [Harten et. al. (1976)]. In reacting flows these overshoots can trigger a chemical reaction and lead to meaningless results.

The energy in the high modes in a neighborhood of a shock is dissipated as heat in the physical system, but not by Eq. (2.6). It would be advantageous to tune or shape the artificial dissipation so it has a much stronger dissipative effect on the high modes than the low and middle modes thus minimizing the damping will be in smooth regions. The middle range frequencies will still be adequately represented to give accurate shock speeds and the dissipation in the high modes will in some sense approximate the physic situtation.

The artificial dissipation can be shaped by using high even order spatial derivatives or using a nonlinear term. Real heat dissipation and viscosity are represented in the equations by second order spatial derivative terms and are not well tuned for numerical calculations. An important feature which must be considered in designing an artificial dissipative term is that one is computing with only a finite number of modes and the true viscous effects usually are dependent on high modes outside the realm of the computation, or the high modes that are poorly represented by the numerical method. By acknowledging that we don't (or can't) approximate the true dissipative effects we are free to shape the artificial dissipation and design it specifically to enhance a numerical method. By allowing nonlinear terms the dissipation can be made to vary from a small amount for weak shocks to a large amount for strong shocks. By using say fourth order derivatives the dissipation will be more dissipative at high frequencies than second order derivative based dissipation and at the same time be less dissipative in the low and middle frequencies.

Energy Cascade

The second major reason for adding artificial dissipation is to solve the energy cascade problem. Typically, energy enters the system at low wave numbers and cascades upward through the high wave numbers where it is eventually dissipated by molecular viscosity and enters the system as heat (Kolmogoroff hypothesis). In numerical calculations the energy spectrum is limited by the number of mesh points. When there is no artificial dissipation in the system the energy cascade backs up at the higher frequencies and shows up in the calculation as high frequency noise or trash. Some of this energy is aliased or reflected back into the lower wave numbers. This closed loop energy cascade can destroy the accuracy in all wave numbers during even moderately short computations.

Gibbs Phenomenon

The third reason for adding artificial dissipation is to compensate for the inexactness of the spatial approximation. These errors are due to approximating a function by an interpolant whose values agree with the function at a discrete set of mesh points. The errors in the interpolant are most severe near discontinuities in the function being approximated. At these points the continuity conditions used to derive the interpolant break down.

Equation (2.6) preserves the positivity of the density of the solution. In general, however, the numerical interpolant does not. Adding artificial dissipation to the numerical approximation damps the high frequencies and helps prevent the loss of positivity.

Dispersion Error

The fourth and fifth reasons for adding artificial dissipation are also to compensate for inaccuracies in the numerical method. Dispersion errors come from the inexactness in both the time and space differencing methods. The dispersion errors due to the different modes of the solution travelling at different and incorrect velocities can accumulate and destroy the accuracy of the computation. This is particularly true for the higher modes even in calculations of flows which should have only smooth solutions. Increasing the accuracy in both the time and space differencing methods will reduce the dispersion in the low and middle frequencies, but not the high modes. It is best to damp these out by some form of artificial dissipation.

Stabilization of Time Integration Methods

The ability of artificial dissipation to stabilize what may otherwise be an unstable time differencing method for Eq. (2.6) lies in the fact that it shifts the spectrum of the spatial operator such that the solution to the modified equation is mathematically and numerically more stable. This is especially true for such standard methods as forward Euler and improved Euler.

Differential Form

For many problems the artificial dissipation inherent to the time integration method is sufficient to compensate for the energy cascade problem and also the entropy production in weak shocks. For strong shocks it is necessary to add significantly more dissipation. The extra dissipation can be added by explicitly adding a dissipative truncation error to Eq. (2.6). This is done in all the shock calculations in Section VII.

The modified equation for these calculations can be written as

$$W_t + F_x = (\Delta x^k d W_x)_x \; , \; k = 1,3 \tag{5.1}$$

where

$$d = \delta \left(\frac{\Delta t}{\Delta x}\right)^{k-1} |\lambda_{max}|^k = \delta \left(\frac{\Delta t}{\Delta x}\right)^{k-1} (|u| + c)^k \tag{5.2}$$

and λ_{max} is the largest characteristic velocity of the system and δ is called the artificial dissipation coefficient where $\delta>0$ if $k=1$ and $\delta<0$ if $k=3$.

It may seem strange at first to use a first or even third order artificial dissipation term with a fourth or sixth order approximation of the derivatives of the flux function. In calculations with strong shock waves there is not always a one-to-one correlation between the formal order of accuracy of a difference scheme and the true accuracy of the calculation. The most reliable and accurate artificial dissipation terms known happen to be of low order and we are stuck with them until better ones can be developed.

VI. TIME DISCRETIZATION

In chosing the "best" numerical method to integrate the Euler Equations through time one has to consider the accuracy, stability, storage requirements, computational complexity and the relative cost of the different methods. These factors are dependent on each other and tradeoffs must be made as to which criteria are more important for a particular problem.

Spectral Analysis

Both the phase and damping errors depend on the spectrum of the differential equation and the time step size. The time step can be varied during the calculation to reduce the numerical integration errors, but the spectrum of the differential equations is determined by the spatial difference operator. A good integration method depends on how accurately it can integrate a particular set of equations. For this reason the spectrum of the spatial difference operator is the most important guide in selecting an efficient numerical method to integrate through time. The spectrum can be determined by analyzing the linearized continuous time - discrete space approximation of the partial differential equation.

Equation (2.6) is solved after adding artificial dissipation and therefore we must analyze a system of the form

$$\rho_t + \rho_x = \delta \Delta x \rho_{xx} \; . \tag{6.1}$$

A semi-discrete approximation of (6.1) results when the spatial derivatives are approximated by finite differences on a mesh of N points. This system can be written in the form

$$y' = Ay + \delta \Delta x B y = Cy = f(y) \tag{6.2}$$

The prime denotes the derivative by y with respect to time and the vector y is an array of the approximate solution at the mesh points.

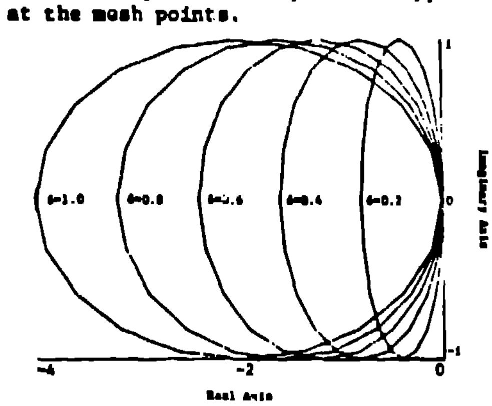

Fig. 3, The eigenvalues of ΔxC

When second-order centered differences are used the eigenvalues of A are imaginary, the eigenvalues of B negative real. The eigenvalues of C are complex and lie on the ellipses graphed in Fig. 3.

We shall first analyze Eq. (6.2) when there is no artificial dissipation (i.e. $\delta=0$) and later include the effects of the dissipation as a perturbation on this equation. When $\delta=0$ Eq. (6.2) is dispersive since the eigenvalues of A lie on the imaginary axis. These eigenvalues, λ, are equal to $i\alpha u$, $i\alpha(u+c)$ and $i\alpha(u-c)$, where α depends upon the spatial order of approximation. When second, fourth or sixth order centered differences are used and the boundary conditions are periodic on the unit interval the corresponding α's are

$$\alpha_2 = (\sin(2\pi j\Delta x))/\Delta x \quad , \tag{6.3}$$

$$\alpha_4 = (8 \sin(2\pi j\Delta x) - \sin(4\pi j\Delta x))/6\Delta x \quad , \tag{6.4}$$

and

$$\alpha_6 = (\sin(6\pi j\Delta x) - 9 \sin(4\pi j\Delta x) + 45 \sin(2\pi j\Delta x))/30\Delta x \tag{6.5}$$

for

$$j = -N/2, -N/2+1, \ldots, N/2 \text{ and } \Delta x = 1/N \ .$$

To facilitate studying the properties of different time integration methods we use a well known result from ordinary differential equations. The isolation theorem, Lomax (1967), states that the stability and accuracy of a numerical integration method for Eq. (6.3) is determined entirely by how it approximates the decoupled diagonalized system

$$y' = \lambda_i y \ , \tag{6.6}$$

with the solution $y_i(t) = y_i(0)e^{\lambda_i t}$ where the λ_i are the eigenvalues of A.

Numerical Methods

Equation (6.6) (hence (6.2)) is a multirate system of equations since some ODE components change on vastly different time scales than others. These systems can have accuracy and stability restrictions that can make standard explicit integration methods inefficient.

We now describe a new class of numerical methods, called iterative multistep (IMS) methods that overcome some of the difficulties in solving multirate systems. These methods are A-Stable when iterated to convergence and converge to the exact solution for linear automous systems of equations.

An example of a common iterative (but not IMS) method is the forward Euler

$$\text{predictor:} \quad y_{n+1}^{(1)} = y_n + \Delta t f_n \ , \tag{6.8}$$

and the improved Euler

$$\text{corrector:} \quad y_{n+1}^{(i)} = y_{n+1}^{(i-1)} + \tfrac{1}{2}\Delta t [f_{n+1}^{(i-1)} - f_n] \tag{6.9}$$

for $i = 2,3,4\ldots$. Here n+1 refers to time t_{n+1} and i is the iteration index.

The regions of absolute stability for this method are symmetric about the real axis and are shown in Fig. 4 , below

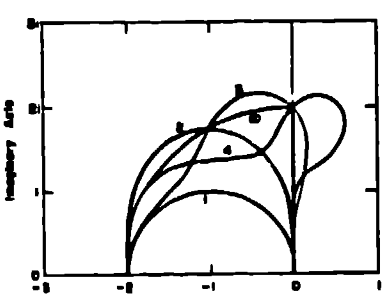

Fig. 4 – Stability regions for (6.8) for $i=1,2,3,4,\infty$.

The method is stable if Δt is chosen small enough that $\lambda\Delta t$ lies within its stability region. For Eq. (6.3) this must hold for all the eigenvalues of C.

The stability of the improved Euler method increases for the first few iterations as seen in Fig. 4 . After three iterations the stability stagnates to the restriction $|\lambda_{max}|\ \Delta t \leq 2$. Also, even when the method does converge, it converges to a solution of the difference equation not the differential equation.

The IMS methods were developed on the premise that if we are willing to do extra work by iterating then it is not unreasonable to expect the stability and accuracy to improve on each and every iteration. These methods are based on the simple recurrence relation

$$y_{n+1}^{(i)} = y_{n+1}^{(i-1)} + c_i \Delta t (f_{n+1}^{(i-1)} - f_{n+1}^{(i-2)}) \tag{6.9}$$

for $i = 3,4,\ldots$. That is, after the corrector cycle a different corrector is used for each additional iteration. The constants c_i depend on the iteration count and the predictor-corrector method used to start the process. The c_i are chosen to increase the order of accuracy of the method for linear automous systems and each iteration. Hence, when iterated an infinite number of times (or to convergence) the method is of infinite order and converges to the exact solution, i.e. the method is A-Stable.

The simplest IMS method, called the iterated Runge-Kutta method, is based on improved Euler where $c_i = 1/i$, $i = 3,4\ldots$. The stability of this method increases with each iteration, as shown in Fig. 5 . For the first four iterations these regions are equivalent to the stability regions of a Runge-Kutta method.

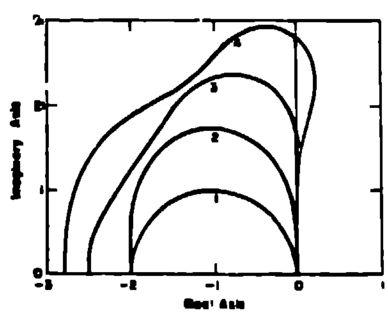

Fig. 5 – Stability regions for the iterated Runge-Kutta method.

Other methods where the coefficients of the IMS methods have been derived are based on the Adams-Bashford-Moulton predictor-corrector sequence, polynomial extrapolation with a backward difference corrector and a new leap-frog predictor-corrector sequence. Each method has a unique corrector sequence and different stability regions. These stability regions can be used as a guide to chose a good method depending on the eigenvalues of the differential equations being solved. For example, the second order leap-frog predictor is given by

$$y_{n+1}^{(1)} = (1-r^2)y_n + r^2 y_{n-1} + \Delta t(1+r)f_n \ , \tag{6.10a}$$

where

$$r = (t_n - t_{n-1})/(t_{n+1} - t_n) \quad ,$$

and the third-order leap-frog corrector is

$$y_{n+1}^{(2)} = [(2-r)(1+r)^2 y_n + r^3 y_{n-1} + \Delta t(1+r)^2 f_n + \Delta t(1+r) f_{n+1}^{(1)}]/(2+3r) \ . \tag{6.10b}$$

The IMS coefficients for this method are c_i = 3/10, 7/30, 4/21, 415/2600, 314/2255, 1153/9420, and 126/1153 for i = 3,4,...9 when r=1. The c_i are functions of r and are not known for general r at this time.

The iterated leap-frog method has stability regions that are particularly good along the imaginary axis, as seen in Fig. 6 below

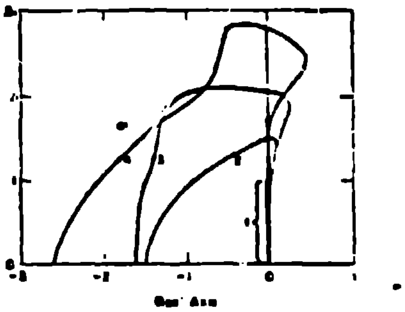

Fig. 6 - Stability regions for the iterated leap-frog when i = 1,2,3 and 4.

The leap-frog predictor is unstable for systems of equations with eigenvalues having a nonzero real part. Therefore, when artificial dissipation is added or the boundary conditions shift the spectrum of the discretized equation the leap-frog method cannot be used without the corrector cycle. The first corrector application extends the bound on the maximum time step by 50%, increases the method to third order and is stable in smooth regions of the solution with or without any spatial artificial dissipation. Another difficulty with using the leap-frog predictor is a unique type of error due to time and space mesh decoupling. The odd and even points of a mesh are only weakly coupled when integrating conservation laws and errors with frequency = 2Δx can degrade the accuracy of the solution with high frequency noise. The corrector cycle couples the mesh points among the three time levels and prevents this weak instability.

A major advantage of the IMS methods is that they allow for local improvements in the stability and accuracy of the calculation. Since only a single time level is used in the iteration only the ODE components that have failed to pass some accuracy test need be iterated on. That is, by iterating locally in regions of rapid changes such as in shock fronts, boundary layers or regions with a refined mesh, the stability and accuracy of the calculation is improved precisely where it is needed. This approach can be used in many problems with severe local stability requirements rather than the more complicated implicit methods or the method of independent timesteps (See Porter (1977)).

When integrating nonlinear equations the IMS methods reduce to the order of the predictor-corrector or Runge-Kutta starting methods. The stability regions still expand with extra iterations but the order of accuracy remains the same. The coefficients for the IMS methods can also be chosen to increase the stability by a maximum amount on each iteration while retaining the order of accuracy of the starting method.

Stability

The numerical solution of (6.8) for an iterative M-step method can be written in the form

$$y(t) = y_0 \sum_{i=1}^{M} \alpha_i e^{\hat{\lambda}_i t} \equiv y_0 e^{\hat{\lambda} t} \qquad (6.11)$$

where $\hat{\lambda} \equiv \lambda$ is the principle eigenvalue of the discretized equation. A numerical method is stable if $Re(\lambda) \le 0$ implies $Re(\hat{\lambda}_i) \le 0$. For most numerical method it is the largest eigenvalue λ_{max} of the linearized system that determines the stability condition.

We now relate the stability diagrams in Figs. (4) - (6) to stability restrictions on Δt when solving the Euler equations. When second-order centered differences are used in space and the leap-frog method is used in time the stability condition is $\Delta t |\lambda_{max}| \le 1$, or

$$\Delta t \max (|u| + c) \left|\frac{\sin(2\pi i \Delta x)}{\Delta x}\right| \le 1$$

or

$$\frac{\Delta t}{\Delta x} \max_{c,u}(|u| + c) \le 1 \qquad (6.12)$$

This is the usual Courant- Friedrichs-Lewy stability condition for explicit methods when solving the Euler equations. If fourth-order centered differences are used in space and the leap-frog predictor-corrector method in time, the corresponding stability condition is

$$\Delta t \frac{8}{6\Delta x} \max (|u| + c) \le 1.5$$

or

$$\frac{\Delta t}{\Delta x} \max (|u| + c) \le 1.125 . \qquad (6.13)$$

Notice in Fig. 4 that some integration schemes such as forward Euler are unconditionally unstable for all Δt > 0 when the spectrum of the discretized system lies on the imaginary axis. It is well known that forward Euler is the heart of many standard methods to solve Eq. (2.6) and in fact is not always unconditionally unstable. This is because of the addition of artificial dissiptation shifts the eigenvalues of the linearized system to the left so they have a negative real part.

We caution the reader that this stability analysis is linear and is not necessarily valid for highly nonlinear phenomenon such as shockwaves. In practice to prevent nonlinear instabilities, it is necessary to restrict the time step such that a shock will not move more than one mesh point per time step.

Accuracy

The accuracy of a method depends on the phase or dispersion error $e_p = Im(\hat{\lambda}-\lambda)$ and the damping or dissipation error $e_d = Re(\hat{\lambda}-\lambda)$. The graphs in Fig. show how these errors are reduced for hyperbolic systems ($Re(\lambda) = 0$) with extra corrector iterations.

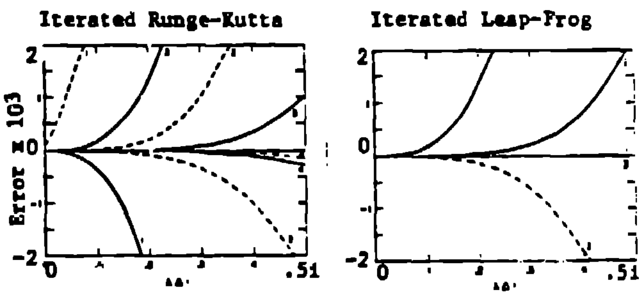

Fig. 7 - Phase error (—) and damping error (---) for the IRK and ILF methods.

These phase and damping error diagrams present much more relevant information than a Taylor series truncation error analysis is capable of. For example, the phase and damping error for the improved Euler method with a single corrector cycle is almost the negative of the phase and damping errors with two corrector cycles. Therefore, we would expect a large increase in accuracy by alternating between one and two correctors on every other time step. This has been found to be the case.

The time step Δt should be chosen to be approximately the same size as the time scale τ for the phenomenon being modeled. The stability restriction for explicit methods requires the time step be proportional to $\lambda_{max} = \max(|u|+c)\Delta x$, the fastest signal speed and shortest time scale $\tau_{min} = 1/\lambda_{max}$ of the system. There are many problems where the important time scale of the problem depends on u not c + u. The optimal time step should be chosen such that the errors in the time integration are approximately the same as the errors in the spatial derivative approximations. Therefore, if $c \gg |u|$ the stability restriction for explicit methods requires that we either take much take much smaller time steps than are needed to balance the space and time truncation errors, or we iterate many times per time step with an iterative multistep method.

In these cases it is often best to use the unconditionally stable implicit method. Two of the best are the second order traizodial rule

$$y_{n+1} = y_n + \frac{h}{2}(f_n + f_{n+1}) \qquad (6.14)$$

and the second order backward difference formula

$$y_{n+1} = [(r+1)^2 - r^2 y_n + \Delta t(r+1) f_{n+1}]/(1+2r) \qquad (6.15)$$

where

$$r = (t_{n+1} - t_n)/(t_n - t_{n-1}) \ .$$

At each time step one must solve a nonlinear system of equations with, say, the multigrid algorithm, see Brandt (1977), or a noniterative direct method such as fractional steps.

Recently Beam and Warming (1978) have compared several implicit methods for the numerical solution of the Euler equations. The methods described in their survey can be implemented using the same techniques for spatial differencing, incorporating boundary conditions and artificial dissipation that are described for explicit methods in this paper.

In calculations when $c < |u|$ then the IMS methods are among the most accurate methods available requiring the least amount of work. The time step may be varied to keep the approximation within a certain accuracy tolerance within the allowable stability restriction. This is easily done with iterative methods by comparing the difference between the predicted value and the first corrected value.

VII NUMERICAL METHODS

Composite Methods

The general flow of a MOL computer code has a well defined structure. The code must:

1. Define the initial conditions for the PDEs.
2. Incorporate the boundary conditions into the discrete system.
3. Evaluate and difference the flux functions.
4. Add artificial dissipation and define W_t.
5. Predict the solution and update the time ($t \to t+\Delta t$) or correct the solution (it is unchanged).
6. Repeat the cycle if the problem is unfinished (go to 2).

In this algorithm four basic decisions must be made in steps 2-5. That determines the algorithm. In step 2 we recommend incorporating the boundary conditions into the discrete system by using fictitious points. This approach can be used with any of the procedures described in Section IV. The extrapolation formula for the fictitious points allows more freedom to include information about the PDEs and the boundary conditions into the difference scheme than does using uncentered differences.

In Step 3 we recommend using second-order differences only in the initial debugging stages of the program and later switching to at least fourth-order. The higher order methods reduce the phase errors and the amount of artificial dissipation needed to stabilize the calculation.

In Step 4 the artificial dissipation term added is crucial to the success of any numerical method for shock calculations. Three of the better numerical implementations of the artificial dissipation, Eq. (5.1), will be described later in this section.

Of the ODE methods used in Step 5 the leap-frog predictor-corrector method has outpreformed any other method we have tested. If the solution can be stored on more than two time levels then the higher order Adams-Bashford-Moulton methods may be more competive. (See Shampine and Gordon (1975)).

Of the above decisions, the choice of a good artificial dissipation term has the most potential for improving a method by tuning it to a particular problem.

Improved Artificial Dissipation

The artificial dissipation term in Eq. (5.1) with $\ell = k = 1$ can be implemented numerically as

$$(\Delta x d W_x)_x = \phi^n_{i+\frac{1}{2}} - \phi^n_{i-\frac{1}{2}} \qquad (7.1)$$

where

$$\phi^n_{i+\frac{1}{2}} = (d^n_{i+1} + d^n_i)(W^n_{i+1} - W^n_i)/(2\Delta x) \quad d^n_i = \epsilon_i(|u|+C)^n_i \ .$$

This first-order form is used in all the calculations in the next section.

There are three basic types of artificial dissipation switches which are natural to use with predictor-corrector methods. The first, is to increase the artificial dissipation coefficient ϵ in regions containing a shock and keep the dissipation small in shockless regions. In one-dimensional calculations when a shock can be detected by a discontinuous negative velocity gradient the artificial dissipation coefficient can be increased accordingly. For example, a possible switch is to replace $\phi^n_{i+\frac{1}{2}}$ in (7.1) by $\alpha\phi^n_{i+\frac{1}{2}}$ where $\alpha = 1$ if $d^n_{i+1} > d^n_i + (\Delta x/3)$ and, say, $\alpha = 1/3$ otherwise. This has proved to be an effective switch in many different calculations.

The second type of switch changes the artificial dissipation coefficient in the predictor and corrector cycles. This is a particularly good switch for the iterated Runge-Kutta method. The forward Euler predictor cycle is less stable (Fig. 3) than the improved Euler corrector method and requires more artificial dissipation to be stable for hyperbolic equations. We, therefore, add a large amount of artificial dissipation in the predictor cycle. In the corrector cycle the artificial dissipation coefficient is reduced, set to zero or even reversed in sign to add antidifusion and counteract the effects of the overly diffused predicted solution.

Boris and Book (1976) have developed similar diffusion/antidiffusion switches in their work on flux corrected transport methods. Their switches are optimized for a particular space and time differencing to maintain the monotonicity of the solution and decrease the phase errors.

The third type switch is designed to prevent contact discontinuities from smearing in long-time or steady-state calculations by artificially compressing them. Harten (1978) has proposed modifying Eq. (2.6) by adding the derivative of an artificial compression function to the right hand side. This function is chosen such that a shock or contact discontinuity for Eq. (2.6) is a shock for the modified equation. That is the contact discontinuity is artificially compressed to reduce numerical smearing.

These improved artificial dissipation strategies all help reduce the numerical errors away from shocks, but no one method stands out as best for all problems. For this reason it is one of the most active areas in developing new methods for the Euler equations.

VIII. NUMERICAL RESULTS

Riemann Problem

An initial value problem for the Euler equations is called a Riemann problem if the initial data consists of two constant states. The initial conditions chosen in this example were also used in a survey article by Sod (1978) to allow comparisons with other difference schemes. Initially a gas (γ=1.4) is separated in a shock tube by a diaphragm at x=0.5. The left and right states of the system are:

$$u(x,0) = 0 \qquad\qquad u(x,0) = 0$$
$$\rho(x,0) = 1,\ 0 < x < \tfrac{1}{2}; \qquad \rho(x,0) = 0.125,\ \tfrac{1}{2} < x < 1$$
$$p(x,0) = 1 \qquad\qquad p(x,0) = 0.1$$

at t=0 the diaphragm is burst and by t=0.25 the gases have developed into a shock wave on the far right, a contact discontinuity near x=0.7 and a rarefaction wave to the left of x=0.5 as seen in Fig. 8.

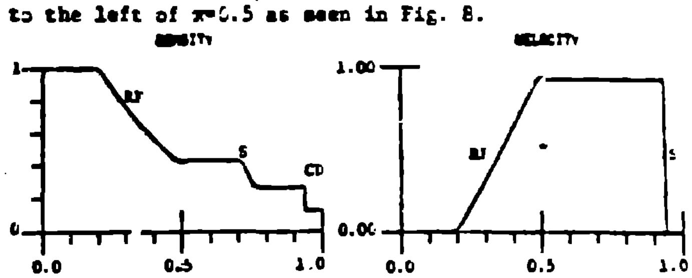

Fig. 8 - Density and velocity at t = 0.25.

The problem was solved on a mesh of 100 points, with fourth order spatial differences and an artificial dissipation term given by Eqns. (5.1) and (7.1) with k=l=1 and δ=0.5. The iterated Runge-Kutta was used in time with one corrector cycle and $|\lambda_{max}\Delta t| = 1$.

When the ends of the shock tube are approximated by a reflecting boundary condition (Eq. (4.4)) the shock reflects from the right boundary and passes through the contact discontinuity by t=0.5 in Fig. 9, below.

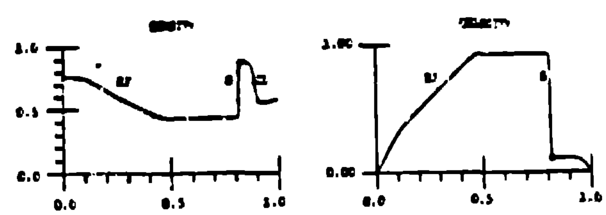

Fig. 9 - Solution t=0.5 with reflecting boundary conditions.

This problem was also solved with a nonreflecting boundary condition at x=1 using the mapping technique given by Eq. (4.17). Three fictitious points were included between x=1 and x=b=1.03 in Fig. 10 below.

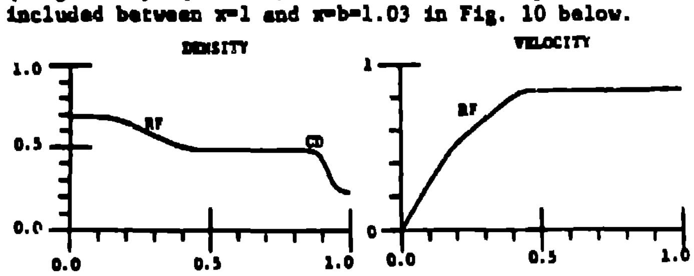

Fig. 10 - Solution at t=0.3 with an artificial boundary at x=1.

Notice the shock has passed through the boundary with almost no reflections.

Generalized Euler Equations

The general one-dimensional form of the Euler equations is

$$(AW)_t + (AF(W))_x = p\begin{pmatrix} 0 \\ A_x \\ -A_t \end{pmatrix} dx, \qquad (8.1)$$

where W and F are as in (2.6) and A(x,t) is an area term depending on the geometry and dimensionality of the problem. For example, A=1 for slab symmetry, A=x for cylindrical symmetry and A=x^2 for spherical symmetry. In this problem A(x,t) is the cross-sectional area of a imploding cylinderical duct.

The cylinder is 100 cm long and collapses at a constant velocity of 1 cm per unit time from a radius of 1 cm to 0.25 cm. The collapse progresses up the cylinder behind a hinge from x=0 at t=0 to x=93.8 at t=7. A shock forms in the gas and is maintained at the hinge location by exponentially accelerating the velocity of the hinge.

This action causes the collapsing cylinder to act as a velocity accelerator. That is, the imploding wall pushes the gas and accelerates it up the cylinder. The velocity of the gas is over 20 times the velocity of the collapsing cylinder walls by the time it has reached the end of the cylinder. A more detailed description of this problem can be found in Colgate et.al. (1977)).

Initially the system is at rest and ρ=0.15, p=1.3 and γ=5/3. The hinge is advanced according to

$$h(t) = 78.93\ \beta\ (\exp(\beta t) - 1)$$

where $\beta = (\gamma-1)/(\gamma+1) = 1/4$. A cross-section of the cylinder, the gas velocity and maximum sound speed (|u|+c) are shown at times t=6 and 7 in Fig. 11

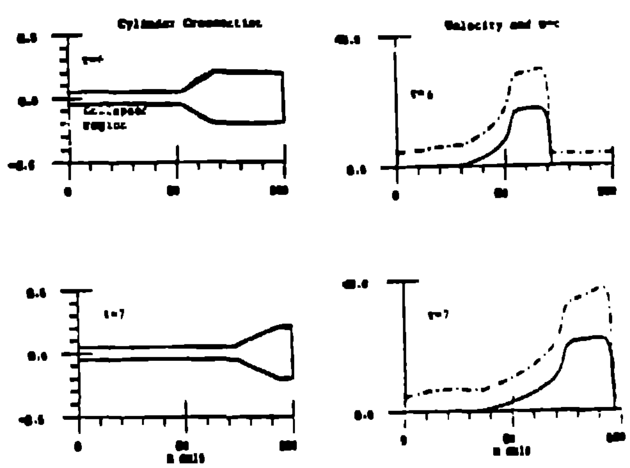

Fig. 11 - The imploding cylinder, the gas velocity (-) and maximum characteristic velocity u+c (---) at times t=6 and 7.

This solution used Eq. (5.1) with Δx=1, δ=0.75, fourth-order differences and the leap-frog predictor-corrector method with $|\lambda_{max}|\Delta t$=1.

IX. SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

In this paper we have followed a MOL approach to constructing accurate and robust numerical methods for hyperbolic PDEs derived from conservation laws. The approach has proved to be straight forward and has led to some excellent new methods for solving the Euler equations. It is our belief that a similar approach may yield some equally useful methods in other "problem areas" of numerical analysis such as the Navier-Stokes equations, reacting or combusting flows and nonlinear diffusion equations.

A major advantage of the MOL approach is the modular structure of the analysis and resulting computer programs. These codes can evolve efficiently since this modularity allows one to test, compare and use the latest numerical methods in the shortest possible time.

Acknowledgements

The author would like to thank Dr. Joe Dendy, Dr. Don Durack and Dr. Burton Wendroff for a number of helpful discussions and suggestions in developing the methods presented in this report.

REFERENCES

Boris, J. P. and Book, D. L. (1976), "Flux-Corrected Transport. III. Minimal-Error FCT Algorithms," J. Comp. Phys. 20, pp. 397–431.

Brandt, A. (1977), "Multi-Level Adaptive Solutions to Boundary-Value Problems," Math. Comp., Vol. 31, No. 138, pp. 333-390.

Colgate, S., Hyman, J. M. and Durack, D. (1977), Los Alamos Scientific Laboratory preprint.

Harten, A., Hyman, J. M., Lax, P. D., and Keyfitz, B. (1976), "On Finite-Difference Approximations and Entropy Conditions for Shocks," Comm. on Pure and Applied Math., Vol. XXIX, pp. 297-322.

Harten A. (1978), "The Artificial Compression Method for Computation of Shocks and Contact Discontinuities: II. Self-Adjusting Hybrid Schemes," Math. of Comp., Vol. 32, 142, pp. 363-389.

Hyman, J. M. (1976), "The Method of Lines Solution of Partial Differential Equations," NYU report COO-3077-139.

Hyman, J. M., (1979), Explicit A-Stable Iterative Methods for the Solution of Differential Equations, in preparation.

Isaacson, E. (1977), "Integration Schemes for Long Term Calculation," Advances in Computer Methods for Partial Differential Equations II, R. Vichnevetsky, ed., pp. 251-255.

Lax, P. D. (1973), Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves, SIAM, Phil., Pa.

Lax, P. D. (1978), The Numerical Solution of the Equations of Fluid Dynamics, Lectures on Combustion Theory, NYU report COO-3077-153, pp. 1-60.

Lomax, H. (1967), An Operational Unification of Finite Difference Methods for the Numerical Integration of Ordinary Differential Equations, NASA TR R-262.

Porter, A. P. (1977), "Independent Timesteps in Numerical Hydrodynamics," UCRL-79608, Rev. 1.

Shampine, L. F. and Gordon, M. K. (1975), Computer Solution of Ordinary Differential Equations - The Initial Value Problem, W. H. Freeman and Company, San Francisco, California.

Sod, G. A. (1977), "A Numerical Study of a Converging Cylindrical Shock," J. Fluid Mech., Vol. 83, 4, pp. 785-794.

Warming, R. F. and Beam, R. M., (1977), "On the Construction and Application of Implicit Factored Schemes for Conservation Laws," SIAM - AMS Proceedings, Vol. 11, Symposium on Computational Fluid Dynamics.